

Weekly Report

April 16, 2017

1 Work

本周报告了一篇使用深度学习处理众包标记的医学影像文章，同时在周五组会向小组同学展示了今年投稿的项目。在投影算法方面，刚开始阅读论文，正在整理思路。

海量高维数据的低维嵌入可视分析

思路主要分为两步：

1. 构造近似knn graph

- 创建多个kd树，划分整个数据集
- 基于kd树，构造一个相对较好的knn graph的初值
- 对于knn graph不断迭代。总体思路是，如果两个数据点同时是另一个数据点的最近邻，那么这两个数据点也有很可能是互为最近邻。通过不断把knn graph扩张，然后对每个节点减去多于k的分支，graph就会靠近最终的knn graph。

2. 低维空间嵌入

- 基于knn graph的结构，我们希望投影到低维空间时，在graph中有更大概率连接的节点相互间靠近，而相距较远的节点尽可能分开，即就是最大化正样本的节点对在kNN图中有连接边的概率，最小化负样本的节点对在kNN图中有连接边的概率（LargeVis）。对于负样本，LargeVis中采用了一些高效的采样技术，可以不用把所有点都计算进来。

李志昊

正在使用doc2vec和lda2vec模型对手机轨迹数据做一些可视化。doc2vec模型使用比较简单，目前正在搭建网页，打算实现二维投影和向量相似度查询。

丁铁成

打算对于出租车轨迹数据使用doc2vec，也许可以在训练过程中加入poi的属性，研究poi和交通路线的关系。本周正在处理轨迹数据，存入数据库。

2 Plan for next week

- 专利
- 专著
- 组会报告

3 Paper Reading

3.1 AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images

因为医学图像标注比较少的背景下，越来越多的工作采用众包进行标注。然而如何统计对同一张图片的多个标记结果成了一个问题。作者采用一个EM算法学习每个众包用户的参数，最后给定一个聚合后的标注结果。实验结果证明，这个方法比少数服从多数的统计结果更好。

3.2 EFANNA: An Extremely Fast Approximate Nearest Neighbor Search Algorithm Based on kNN Graph

本文介绍了一种快速构建近似knn graph的方法，主要包括构建kd树（每个叶子节点可以包含10个数据点）和构造knn graph的初始值（提供一个好的初始值可以降低后面迭代算法的时间）。

3.3 ACTIVIS: Visual Exploration of Industry-Scale Deep Neural Network Models

在预测任务，深度学习已经实现了非常高的准确率，然而深度学习模型对于我们来说还是一个黑盒。本文设计了ACTIVIS系统，用于探索深度网络中每个神经元对于预测结果的影响。

3.4 Visualizing Large-scale and High-dimensional Data

看了LargeVis的文章，将高维向量嵌入到低维空间的降维方法主要包括两个步骤：构建knn graph和把graph嵌入到低维空间。

3.5 node2vec: Scalable Feature Learning for Networks

通过控制在网络中随机游走的方向，可以学习到不同的网络结构。